

Asymptotically optimal model selection method with right censored outcomes

SÜNDÜZ KELEŞ*, MARK VAN DER LAAN** and SANDRINE DUDOIT†

*Division of Biostatistics, University of California, Berkeley CA 94720, USA. E-mail:
*sunduz@stat.berkeley.edu; **laan@stat.berkeley.edu; †sandrine@stat.berkeley.edu*

Over the last two decades, nonparametric and semi-parametric approaches that adapt well-known techniques such as regression methods to the analysis of right censored data, e.g. right censored survival data, have become popular in the statistics literature. However, the problem of choosing the best model (predictor) among a set of proposed models in the right censored data setting has received little attention. We develop a new cross-validation-based model selection method to select among predictors of right censored outcomes such as survival times. The proposed method considers the risk of a given predictor based on the training sample as a parameter of the full data distribution in a right censored data model. Then, the doubly robust locally efficient estimation method or an *ad hoc* inverse probability of censoring weighting method, as presented by Robins and Rotnitzky and later by van der Laan and Robins, is used to estimate this conditional risk parameter based on the validation sample. We prove that, under general conditions, the proposed cross-validated selector is asymptotically equivalent to an oracle benchmark selector based on the true data generating distribution. The method presented covers model selection with right censored data in prediction (univariate and multivariate) and density/hazard estimation problems.

Keywords: cross-validation; density/hazard estimator selection; model selection; multivariate prediction; nonparametric/semi-parametric regression; prediction of survival; right censored data; univariate prediction